2018-01-01

Initial training with difficult items does not facilitate category learning

Edmunds, CER

10.1080/17470218.2017.1370477
Quarterly Journal of Experimental Psychology
SAGE Publications

ARTICLE

Initial training with difficult items does not facilitate category learning

C. E. R. Edmunds[a], Andy J. Wills[a] and Fraser Milton[b]

[a]Plymouth University, UK; [b]University of Exeter, UK.

ARTICLE HISTORY
Compiled July 26, 2017

ABSTRACT
In the phenomenon of transfer along a continuum (TAC), initial training on easy items facilitates later learning of a harder discrimination. TAC is a widely replicated cross-species phenomenon that is well predicted by certain kinds of associative theory (e.g., Sutherland & Mackintosh, 1971). A recent report of an approximately-opposite phenomenon (i.e. facilitation by initial training on hard items, Spiering & Ashby, 2008) poses a puzzle for such theories, but is predicted by a dual-system model (COVIS; Ashby et al., 1998). However, across four experiments we present substantial evidence that Spiering and Ashby's conclusions were in error. Their result appears to be a false positive and, as such, should not form part of the evidence base for COVIS, nor be considered as a counter-example to the pervasive TAC phenomenon.

KEYWORDS
categorisation; transfer along a continuum; implicit; explicit

In the phenomenon of transfer along a continuum (TAC), initial training on easy items facilitates later learning of a harder discrimination (Lawrence, 1952). For example, consider an experiment that aims to teach animals to discriminate between two grey squares that vary slightly in brightness. TAC is shown if the group of animals who were initially trained to discriminate a black square and a white square perform better at test on the grey squares than the animals who were trained on the grey squares throughout (Lawrence, 1952). In humans, TAC has been found across a variety of stimulus types, including faces (Suret & McLaren, 2003), mammograms (Hornsby & Love, 2014) and birdsong (Church, Mercado, Wisniewski, & Liu, 2013).

In a classic experiment, Mackintosh and Little (1970) demonstrated that the TAC effect persists even when the response mappings are reversed between the training and test phase. Their demonstration was with pigeons, but an analogous result was subsequently shown in humans (Suret & McLaren, 2003). The fact that TAC persists across a reversal is generally considered to support the idea that TAC is a consequence of increased attention to the relevant stimulus dimension (Lawrence, 1952; Sutherland & Mackintosh, 1971). Specifically, the participants who were initially trained on the easy discrimination were more easily able to identify and attend to the critical dimension of variability, thus improving their performance on the more difficult task. The

C. E. R. Edmunds. Email: ceredmunds@gmail.com. We would like to acknowledge the contribution to the study of Transfer Along a Continuum made by Professor Nick Mackintosh, FRS, who passed away during the preparation of this article. He will be sadly missed. We also thank Gemma Williams for her assistance in the coding of the strategy-report questionnaires discussed in ths article.

advantage this attentional re-allocation provides is assumed to more than offset the cost introduced by the reversal. Although this account can be expressed in terms of stimulus dimensions, it can also be captured by formal models that are purely elemental in nature (McLaren & Mackintosh, 2002; Suret & McLaren, 2003). Thus, the core of the standard account of TAC is that attention is directed towards those aspects of the stimulus that are the best predictors of the outcome (Mackintosh, 1975). This idea has subsequently motivated a wide range of research on the relation between attention and learning; for a recent review, see Le Pelley, Mitchell, Beesley, George, and Wills (2016).

In contrast to the empirical and theoretical consensus concerning TAC, Spiering and Ashby (2008) reported an approximately opposite finding. In their experiment, participants were trained on a two-category discrimination task using one of three training orders: easy-to-hard, hard-to-easy, or random. For brevity, we will focus on the first two orders, as the random condition adds nothing of consequence to the conclusions. In the easy-to-hard condition, participants first saw the easy stimuli, which were furthest from the category boundary (see Figure 1), followed by the moderate-difficulty stimuli, and then the hard stimuli, which were closest to the category boundary. In the hard-to-easy condition, participants saw the stimulus types in the opposite order. Finally, the participants were tested on all the stimuli in the category structure. Spiering and Ashby found that, for the particular category structure shown in Figure 1, participants who were given hard-to-easy training had better performance on the final all-items test than those given easy-to-hard training.

Spiering and Ashby (2008) argue that their result is best explained in terms of the COVIS model (COmpetition between Verbal and Implicit Systems; Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby, Paul, & Maddox, 2011). This model assumes that learning, at least in humans, is mediated by two parallel competing systems. The Explicit System learns by testing verbalisable hypotheses. In contrast, the Procedural system is hypothesised to be implicit and gradually associates stimuli with responses.

According to COVIS, normal adults begin learning using simple verbalisable rules mediated by the Explicit System, only switching to the Procedural System if these rules result in poor performance. For the particular category structure used in their experiment (see Figure 1), a simple rule-based strategy based on a single stimulus dimension (e.g., bar frequency) results in excellent performance with the easy stimuli (100%), but poor performance when classifying the moderate (80%) and difficult (60%) stimuli closer to the category boundary.[1] The Procedural System, given sufficient opportunity, can perform well on all three stimulus types. Hence, according to COVIS, participants in both conditions would tend to eventually switch to the Procedural System, but would do so at different times. Participants in the easy-to-hard condition would switch to the Procedural System relatively late in training because using the Explicit System initially resulted in high levels of accuracy. In contrast, the participants in the hard-to-easy condition would score lower using a simple unidimensional rule from the outset, and so would more quickly realise that "no explicit strategies will succeed" (Spiering & Ashby, 2008, p. 1171). They would then more quickly switch to using the more optimal Procedural System to learn the structure, and so score more highly on the moderate difficulty (Block 2) items, and in the final all-items test (Block 4).

The difference in final test performance found by Spiering and Ashby (2008) can be interpreted as consistent with Spiering and Ashby's account, or with the

[1]Note that Figure 1 slightly misrepresents these accuracy scores, due to drawing the stimuli across areas of space rather than as points.

approximately-opposite TAC result, depending on whether one assumes it is the initial (Block 1) training, or the most recent (Block 3) training that primarily determines performance on the final test. Spiering and Ashby argue that the effect of training order they observe is due to the type of stimuli the participant saw initially (in Block 1). However, the stimuli in the training block just prior to the final test block (Block 3) are hard in the easy-to-hard condition and easy in the hard-to-easy condition. Therefore, it is possible that participants' performance in the final all-items test phase reflects the most recent, rather than the initial, training. If this were the case, their results would be consistent with TAC: participants who saw the easy stimuli in Block 3 performed better in Block 4 than those who saw the hard stimuli in Block 3.

However, Spiering and Ashby's experiment contains a further result that more directly supports their conclusion. Specifically, performance in Block 2 (moderate difficulty items) was better if Block 1 contained hard stimuli than if Block 1 contained easy stimuli. This directly supports their conclusion that initial training on difficult items improves performance, relative to initial training on easy items. This result, being approximately opposite to TAC, poses something of a puzzle for theories, such as Sutherland and Mackintosh (1971), that predict the presence of TAC effects.

Experiment 1

In Experiment 1, we attempted to directly replicate the first two blocks of Experiment 1 of Spiering and Ashby (2008). We did not include the final two blocks, due to the confound inherent in that part of their design (see the above discussion of the initial-training vs. recent-training confound), but our experiment was otherwise identical to theirs. We expected this attempt at replication to be successful, on the grounds that previous published attempts to directly replicate parts of experiments from Ashby's lab have been successful. For example, previous critiques of COVIS-inspired empirical work have primarily been based on the presence of confounds, rather than failure to replicate *per se* (Edmunds, Milton, & Wills, 2015; Newell, Dunn, & Kalish, 2010; Newell, Moore, Wills, & Milton, 2013).

Method

Participants

The participants were 40 undergraduate psychology students recruited from Plymouth University participant pool and were randomly assigned to one of the two conditions (N=20 in each). They received research credit in exchange for their participation.

Category structure and stimuli

The stimuli were sine-wave gratings displayed on a grey background, and were identical to those used in Spiering and Ashby (2008). The stimuli are shown in Figure 1.

Design

This experiment had a single between-subjects factor with two levels: easy-to-moderate and hard-to-moderate. In Block 1, participants in the easy-to-moderate condition were shown stimuli that were easy to classify (as they were far from the category boundary).

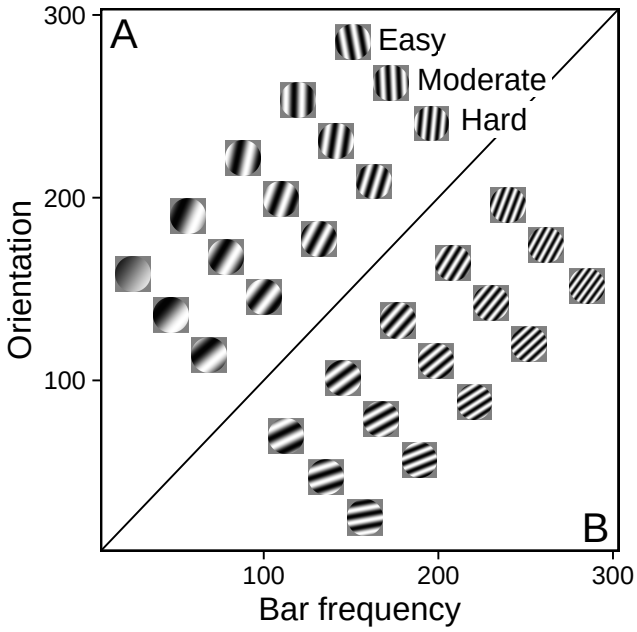

Figure 1. The stimuli used in Experiment 1 represented in abstract stimulus space. The diagonal line represents the optimal category bound. The two categories are labelled A and B. Also included are the stimulus difficulties for Category A: the stimuli furthest from the decision boundary are the easy stimuli and those closest to the category boundary are the hard stimuli.

In contrast, participants in the hard-to-moderate condition were shown stimuli that were difficult to classify (as they were close to the decision boundary). Then in Block 2, participants in both conditions were shown stimuli that were moderately difficult to classify. The design was identical to the first two blocks of Spiering and Ashby (2008). However, unlike Spiering and Ashby, we did not include a random condition, as this condition appeared to provide no useful additional information.

Materials

The experiment was run using MATLAB with the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) extensions on a desktop computer with a 21.5-inch screen.

Procedure

On each trial, a stimulus and the category labels were displayed on a white background until the participant had responded by pressing either the "D" or "L" key. If a participant failed to respond after 5000ms had passed, a screen displaying "PLEASE RESPOND FASTER" was shown to them for 500ms. If the participant responded, 500ms of audio feedback was played to them over headphones. For correct responses, the tone was a 262Hz sine-wave, which sounds similar to a low-pitch tuning fork. For incorrect responses, the tone was a 400Hz saw-tooth, which is a higher-pitched, harsher sound. The inter-trial interval was 1500ms.

There were two blocks of 150 trials each, resulting in a total of 300 trials. In each block, 10 stimuli were presented in a random order 15 times. The stimuli presented in each block depended on the condition to which the participant was assigned. In Block 1, participants in the easy-to-moderate condition were shown only the easy stimuli, those

furthest from the decision boundary; participants in the hard-to-moderate condition were shown only the hard stimuli, those closest to the decision boundary. In Block 2, the participants in both conditions were shown the moderately difficult stimuli in a random order, 15 times each, with feedback.

Additionally, after the experiment was completed, participants were asked to fill in a questionnaire. This aimed to determine which strategy they had used to categorise the stimuli. They were asked to "Imagine that another person was asked to complete the experiment as you did. What instructions would you give them so that they could exactly copy your pattern of responding?" They were given a large box in which to fill in their answer and asked to respond as precisely as possible.

Analysis

We calculated Bayes Factors. This is because, in traditional null-hypothesis significance testing, non-significant results are ambiguous: they could either be due to insufficient statistical power or due to the null hypothesis being correct (Dienes, 2011). It is important to be able to distinguish between these two possibilities.

By convention, if the Bayes Factor is over three then the experiment has found evidence for the experimental hypothesis, whereas if the Bayes Factor is less than a third, the experiment finds evidence for the null hypothesis (Jeffreys, 1961). A Bayes Factor of one indicates that the evidence is exactly neutral with respect to the experimental and null hypotheses (Dienes, 2011). Values between a third and three are typically interpreted as indicating that the experiment was not sensitive enough and no conclusions can be drawn.

The Bayes Factors for the accuracy data in the experiments in this article were calculated according to the procedure recommended by Dienes (2011) using the R script implemented by Baguley and Kaye (2010). The predicted differences between the easy-to-moderate and hard-to-moderate conditions were estimated directly from Spiering and Ashby (2008). The inclusion of the mean difference in Spiering and Ashby's original study as the prior of our Bayes Factor calculation treats their result as the evidence available prior to running our studies, and then uses the result of our studies to update that evidence.

For Block 1, we assumed a two-tailed normal distribution for the prior with a predicted mean difference of 0.101 and predicted standard deviation of 0.054. [2] For Block 2, we assumed a two-tailed normal distribution with predicted mean difference of -0.139 and standard deviation of 0.07. Bayes Factors for the reaction time data were not calculated as average reaction times were not reported in the original Spiering and Ashby (2008) experiments.

All data analyses were conducted in R (R Core Team, 2017). All trials for which the reaction time was greater than 5000ms were removed.

Results

Three participants in the easy-to-moderate condition scored considerably below chance (less than 0.15 in Block 1 compared to chance at 0.50) and so were excluded from the following analyses. The results are displayed in Figure 2. The trial-level raw data

[2] In Dienes (2011), the standard deviation of the prior is typically defined as half the mean; this captures the belief that the true mean difference could plausibly take a range of values, but that an effect in the opposite direction to that previously observed is unlikely. A side effect is that the prior for small effects is more precise (lower s.d.) than the prior for large effects.

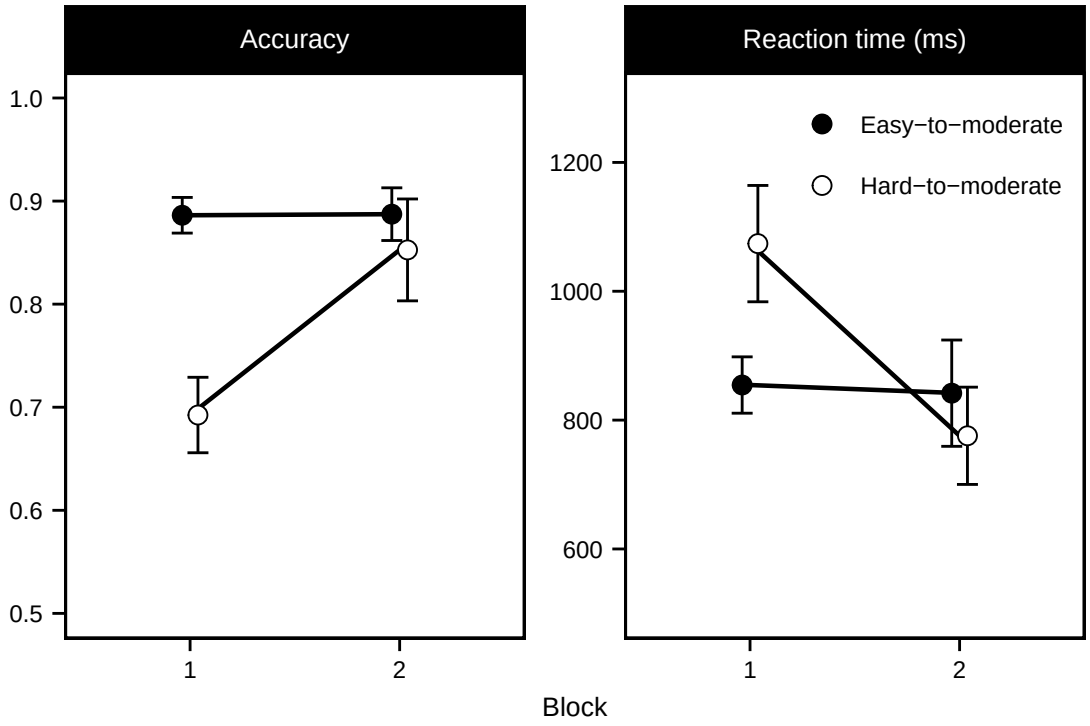

Figure 2. The average accuracy and reaction time for each block in each condition in Experiment 1. Error bars are difference-adjusted between-subject 95% confidence intervals (Baguley, 2012).

for Experiment 1 are available at `www.willslab.org.uk/ply22` with md5 checksum `69e146c069d338bd4e2d2c0029bda3dd`.[3]

During Block 1, as expected, performance on the hard stimuli was worse than performance on the easy stimuli, $t(35) = 6.71$, $d = 2.21$, $p < .001$, $BF > 1 \times 10^6$. However, this initial difference in accuracy had no effect on accuracy in Block 2, $t(35) = 0.88$, $d = 0.29$, $p = .387$. Indeed, there was substantial evidence for the null as the Bayes Factor was less than a third, $BF = 0.07$. The Bayes Factor remains in favour of the null hypothesis ($BF < 1/3$) even if the predicted mean difference between the conditions is reduced by two thirds to -0.046, $SD_{\text{diff}} = 0.023$.

The Block 1 sample mean difference between conditions was 0.194, with a standard error of 0.029. The Block 2 sample mean difference was 0.035, with a standard error of 0.040.

The reaction time data was consistent with the accuracy data. During Block 1, reaction time for the hard stimuli was higher than reaction time for the easy stimuli, $t(35) = 3.07$, $d = 1.01$, $p = .004$. However, in Block 2 there was no significant difference between conditions, $t(35) = 0.88$, $d = 0.29$, $p = .384$. Additional strategy analyses are reported towards the end of the paper.

[3]Publication of a checksum allows the reader to independently confirm that the raw data in the archive is unchanged.

Discussion

Using the category structure illustrated in Figure 1, Spiering and Ashby (2008) found that hard initial training improved performance on moderate-difficulty stimuli, relative to easy initial training. They argued that this finding supported the COVIS model of category learning (Ashby et al., 1998, 2011). Their result is also an approximately opposite effect to the well-established phenomenon of Transfer Along a Continuum (Lawrence, 1952). In Experiment 1, we aimed to replicate Spiering and Ashby's effect. However, in contrast to Spiering and Ashby, we failed to find an advantage for either training order. Indeed, we found substantial Bayesian evidence for the null hypothesis. In other words, initially seeing the easy or hard stimuli had no effect on performance with the subsequent moderate-difficulty stimuli.

Experiment 2

It is possible that we failed to find an effect in Experiment 1 because certain aspects of the procedure added additional noise, thereby obscuring the effect. First, it is plausible that the unusual choice of feedback could have misled some of our participants. In both Spiering and Ashby (2008) and Experiment 1, the feedback was a 500ms tone administered over headphones. However, the mapping of tone pitch to the feedback was not intuitive: the higher tone indicated incorrect responses and the lower tone indicated correct responses. This is not common practice in other studies within the COVIS canon (e.g., Ell & Ashby, 2006), nor more broadly in experimental psychology, or even in other non-experimental settings such as game shows. So it is possible that this feedback may have been consistently misinterpreted by some participants. This idea is further supported by the fact that three participants had scores well below chance (<15%). This level of performance indicates that they learned the category structure but pressed the wrong keys for each category. Furthermore, as Spiering and Ashby did not apply a learning criterion, it is possible that their sample also included participants like these, that went undetected in their analyses. Failing to use learning criteria has previously produced interpretative difficulties in some other COVIS-inspired experiments (e.g., Newell et al., 2010).[4]

Another feature of the procedure that may have added additional noise is the choice of stimuli. In the verbal reports of Experiment 1, which are discussed in more detail in a later section, participants described several eclectic stimulus features that appeared to map onto the bar width dimension. These features included how "zoomed in" the stimulus was, whether the stimulus was symmetrical or not, how many bars there were, and the amount of contrast between light and dark. Using these representations may undermine the inferences we wish to draw from this experiment. For example, it is possible that the "zoom" property maps onto the experimenter-defined dimension of interest (bar width) in a non-linear way. This type of mapping corresponds to sub-regions of the stimulus space being stretched, which may alter the representation of the category structure in ways that are hard to predict.

To address these potential issues, in Experiment 2, we repeated Experiment 1 using line stimuli varying in length and angle, and visual feedback, i.e. "Correct" or "Incorrect!". Line stimuli have been used in previous COVIS-inspired experiments (e.g., Filoteo, Lauritzen, & Maddox, 2010) and seem less likely to produce eclectic stimulus

[4]We requested Spiering and Ashby's trial-level raw data but did not receive a data set that they wanted to endorse as veridical.

representations than Experiment 1's sine-wave gratings. Visual feedback is commonplace throughout the study of category learning, and seemed likely to be less confusing than Experiment 1's counter-intuitive tone-based feedback.

Method

Participants

The participants were 43 undergraduate psychology students recruited from the Plymouth University participation pool. They were randomly assigned to either the easy-to-moderate condition (N=20) or the hard-to-moderate condition (N=23). They received research credit in exchange for their participation.

Category structure and stimuli

The abstract category structure was identical to that used in Experiment 1 of Spiering and Ashby (2008) and Experiment 1 above. However, this category structure was instantiated with black line stimuli that appeared on a white background. These stimuli varied in the length of the line and its orientation. The variation in the length of the lines were matched to the variation of line length in previous research (e.g., Edmunds et al., 2015; Filoteo et al., 2010). To do this, we calculated the linear scaling factor that would transform the bar-frequency value to a corresponding line-length value, such that the minimum and maximum values were 25 and 285 respectively. These values were the length of the line in pixels. The orientation of the lines was the same as the orientation of the sine-wave gratings in Experiment 1.

Procedure

The procedure for this experiment was identical to that of Experiment 1 aside from changing the feedback type. Rather than using 500ms tones, we displayed either 'Correct' or 'Incorrect!' in black in the centre of the screen for 500ms. Also, due to an oversight, we did not give these participants the strategy questionnaire after training.

Analysis

All trials for which the reaction time was greater than 5000ms were removed.

Results

No participants scored below chance, so the following analyses are conducted on all participants. The results are displayed in Figure 3. The trial-level raw data are available at `www.willslab.org.uk/ply12` with md5 checksum `3e3f1fb62d3d810b201b6b1c00f1fb86`.

During Block 1, as expected, performance on the hard stimuli was indeed worse than performance on the easy stimuli, $t(41) = 7.52$, $d = 2.30$, $p < .001$, $BF > 1 \times 10^6$. However, this initial difference in accuracy had no effect on learning performance in Block 2, $t(41) = 1.18$, $d = 0.36$, $p = .246$. Indeed, there was substantial evidence for the null, $BF = 0.05$. The Bayes Factor remains in favour of the null hypothesis ($BF < 1/3$) even if the predicted mean difference is reduced by 6/7 to -0.02, $SD_{\text{diff}} = 0.01$. The Bayes Factors were calculated using the same technique and prior as described in

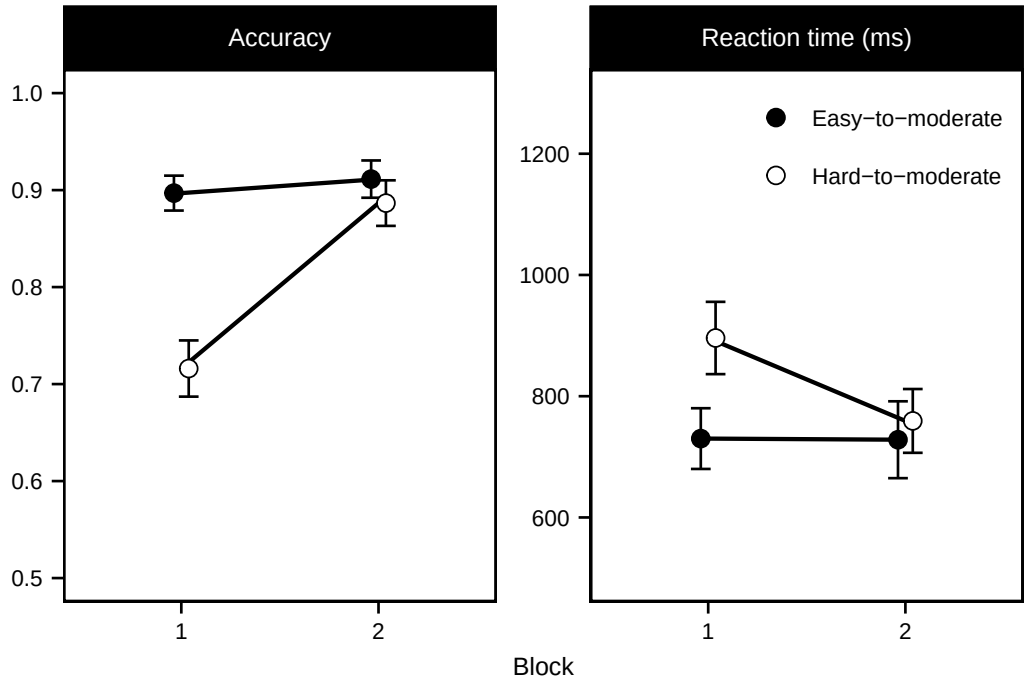

Figure 3. The average accuracy and reaction time for each block in each condition in Experiment 2. Error bars are difference-adjusted between-subject 95% confidence intervals (Baguley, 2012).

Experiment 1. Here, in Block 1, the sample mean difference was 0.181, with a standard error of 0.024. For Block 2, the sample mean difference was 0.025, with a sample standard deviation of the difference of 0.021.

The reaction time data was consistent with the accuracy data. During Block 1, performance on the hard stimuli was slower than responding on the easy stimuli, $t(41) = 2.96$, $d = 0.90$, $p = .005$. However, in Block 2 the difference between conditions was not significant, $t(41) = 0.58$, $d = 0.18$, $p = .567$.

Discussion

Spiering and Ashby (2008) found that participants who were initially trained on a difficult discrimination had better performance on subsequent moderate-difficulty items than participants who were initially trained on the easy version of that discrimination. In contrast, in an attempted replication reported in Experiment 1, we found evidence for the absence of a difference between training-order conditions. We postulated that this may have been due to the non-intuitive choice of feedback: the mapping from correct/incorrect to tone pitch was opposite to that usually seen in psychology experiments. We also thought that the psychological stimulus representation of the sine-wave gratings of Experiment 1 might be more complex and varied than the experimenter-defined representation. Both these things might, at the very least, add noise and obscure the presence of an effect. In Experiment 2, we attempted to address these potential issues by changing the feedback from tones to visual feedback, and using lines varying in length and angle, rather than sine-wave gratings. However, this failed to make a difference: participants in both conditions still performed equally well in the final block.

Experiment 3

In Experiments 1 and 2, we examined whether hard-to-moderate training, compared to easy-to-moderate training, resulted in superior performance on learning the category structure shown in Figure 1. Contrary to Spiering and Ashby (2008), we failed to find an effect of initial training type on final performance: both easy and hard initial training resulted in the same level of performance in Block 2. Additionally, a Bayesian analysis found evidence for the null hypothesis. This suggests that there is genuinely no difference between our training-order conditions. It also suggests that the effect reported by Spiering and Ashby may have been a false positive.

That being said, Spiering and Ashby (2008) found a difference at two points in their four-block design: in Block 2 with just the moderately-difficult stimuli, and also in Block 4 with all the stimuli. Our Experiments 1 and 2 only examined participants' performance on the moderately-difficult stimuli. Therefore, it may be that if we had included all the stimuli at test we might have found an effect of training order. Experiment 3 therefore employs an all-items test, rather than a moderate-difficulty-items test, to examine this possibility.

Aside from the absence of a training-order effect, the results of our Experiments 1 and 2 differ from those of Spiering and Ashby (2008) in another way. Specifically, performance in Block 2 of their experiment (averaged across conditions) was about 80%. In both of our experiments, it was closer to 90%. Hence, for whatever reason, it seems our participants did a bit better on this task than Spiering and Ashby's participants. Thus, one possible explanation of why Spiering and Ashby observed a training-order effect whilst we did not is that, in our studies, it is obscured by a ceiling effect. In Experiment 3, we investigated this possibility by reducing stimulus presentation time, which (we presumed) would reduce overall performance levels.[5]

Method

Participants

The participants were 38 undergraduate psychology students recruited from the Plymouth University participation pool. They were randomly assigned to either the easy-to-all (N=18) or hard-to-all (N=20) conditions. They were awarded research credit in exchange for their participation.

Procedure

The format of the experiment remained similar to that in Experiment 2. However, now the line stimulus was shown for 350ms rather than until the participant responded. Additionally, after the stimulus, a mask was displayed until the participant responded. The mask was constructed by placing every line stimulus three times on a white background with each end of the stimulus randomly displaced along both stimulus dimensions by a number of pixels randomly drawn from a uniform distribution between -120 and 120. The mask ensured that participants could only visually process the to-be-classified stimulus for the allotted 350ms.

As in Experiment 1, at the end of the experiment participants were also asked to

[5]We initially thought that just changing from a moderate-difficulty test to an all-stimuli test might, by itself, lower overall Block 2 performance. However, a pilot study, not reported in this paper, discounted this possibility; the pattern of performance was almost identical to that in the previous experiments.

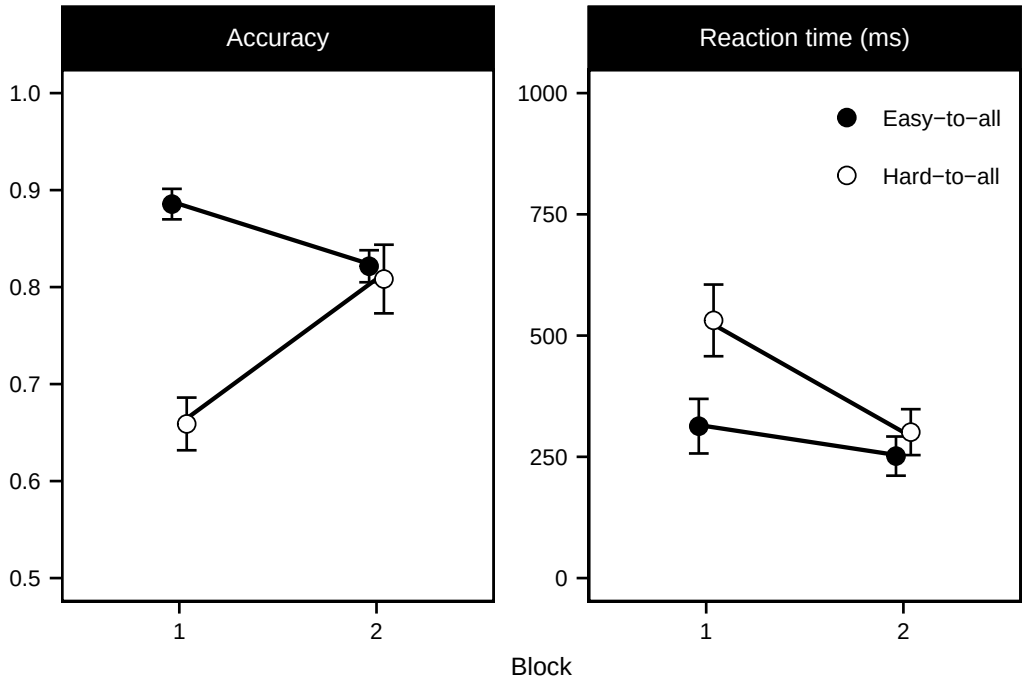

Figure 4. The average accuracy and reaction time for each block in each condition in Experiment 3. Error bars are difference-adjusted between-subject 95% confidence intervals (Baguley, 2012).

report the strategy that they used to learn the category structure.

Analysis

All trials for which the reaction time was greater than 5000ms were removed.

Results

No participants scored below chance, so the following analyses were conducted on all participants. The results are displayed in Figure 4. The trial-level raw data are available at `www.willslab.org.uk/ply42` with md5 checksum `ccec6982393ac850f7f04f87107210d`.

During Block 1, as expected, performance on the hard stimuli was worse than performance on the easy stimuli, $t(36) = 10.40$, $d = 3.38$, $p < .001$, $BF > 1 \times 10^6$. The reduction in stimulus presentation time, relative to Experiments 1–2, also had the expected effect, with mean Block 2 performance (across conditions) close to 80%. However, the Block 1 difference in accuracy still had no effect on learning performance in Block 2, $t(36) = 0.48$, $d = 0.157$, $p = .633$. Indeed, there was substantial evidence for the null, $BF = 0.05$. The Bayes Factor remains in favour of the null hypothesis ($BF < 1/3$) even if the predicted mean difference is reduced by 3/4, to -0.03475, $SD_{\text{diff}} = 0.0175$.

These Bayes Factors were calculated using the same technique and prior as described in Experiment 1. Here, for Block 1, the sample mean difference was 0.227, with a sample standard error of 0.022. For Block 2, the sample mean difference was 0.013, with a sample standard error of 0.027. As the current experiment looked at performance on

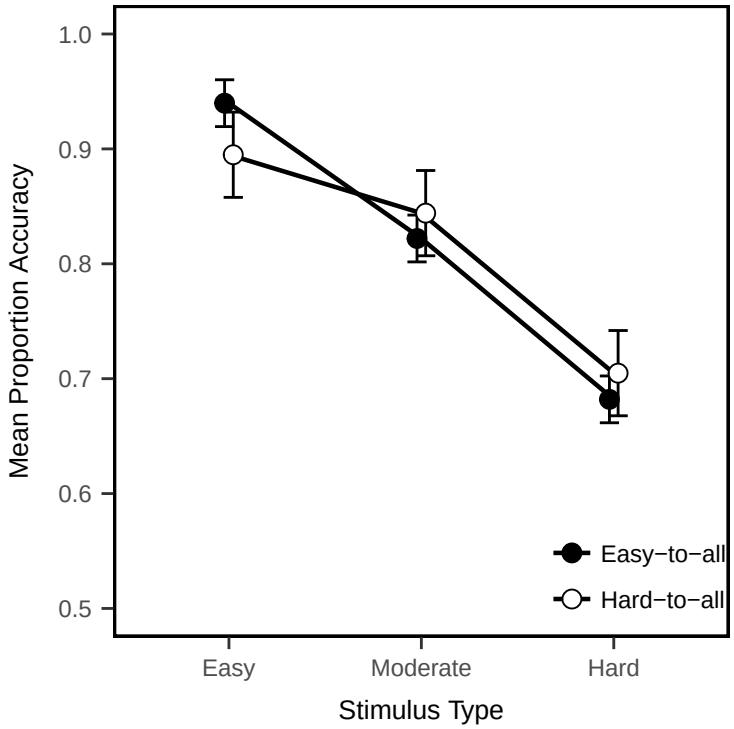

Figure 5. The average accuracy for each stimulus difficulty level in Experiment 3. Error bars are difference-adjusted between-subject 95% confidence intervals (Baguley, 2012).

all stimuli (easy, moderate, and hard) in Block 2, some might argue that using a prior based on Block 4 of Experiment 1 of Spiering and Ashby (2008) might be more appropriate (as this was the block in their experiment in which all stimulus difficulties were presented). Use of this prior makes no difference to the conclusions drawn about Experiment 3.

The reaction time data was consistent with the accuracy data. During Block 1, responding on the hard stimuli was slower than responding on the easy stimuli, $t(36) = 3.42$, $d = 1.11$, $p = .002$. In Block 2, the difference between conditions was not significant, $t(36) = 1.16$, $d = 0.378$, $p = .252$.

Additionally, because the participants in each condition saw all the stimuli in Block 2, we were able to examine the difference between the conditions at each level of stimulus difficulty (see Figure 5). A mixed ANOVA found a significant main effect of stimulus difficulty, $F(2, 72) = 167.76$, $\eta^2_G = 0.50$, $p < .001$. As stimulus difficulty increases, average accuracy decreases. Additionally, there was a significant interaction between stimulus difficulty and condition, $F(2, 72) = 4.95$, $\eta^2_G = 0.03$, $p < .01$. A simple main effects analysis found that the difference between conditions for the easy stimuli approached significance, $t(36) = 1.88$, $p = .068$; performance was numerically higher in the easy-to-all condition than in the hard-to-all condition for these stimuli. The difference between conditions did not reach significance for either the moderate, $t(36) = 0.29$, $p = .772$, or the hard stimuli, $t(36) = 0.32$, $p = .754$.

Discussion

Spiering and Ashby (2008) found that participants who were initially trained on a harder version of a classification task had superior performance to participants who were initially trained on the easiest version of the task. In contrast, Experiment 3, like Experiments 1 and 2, found substantial evidence for the null—the difficulty of initial training did not affect performance on a subsequent test. Experiment 3 added to the previous two demonstrations of a null effect by showing it under conditions where a ceiling effect at test was unlikely (because overall performance at test, averaged across conditions, was around 80% correct, well below ceiling for this task, and similar to the overall performance levels observed in Spiering and Ashby).

Experiment 3 also demonstrated that the null effect was not limited to the moderate-difficulty test items used in Experiments 1–2, but also persisted when the test items were a mix of low-, moderate-, and high-difficulty stimuli. As all the participants saw all the stimuli in the Experiment 3 test phase, we could also look to see whether there was any interaction between stimulus difficulty and condition. Such an interaction was observed, but the simple effects were inconclusive—no significant effect of initial training was observed at any of the three stimulus difficulties.

On the basis of the non-significant trends, one might argue that something akin to a TAC effect was observed, given that easy initial training numerically facilitated test performance on easy items, relative to hard initial training. One argument that this was not an example of TAC comes from the fact that participants in the easy-to-hard condition had seen the easy stimuli before, so this effect may just be an effect of stimulus familiarity. However, speaking against this, there was no such advantage for the hard stimuli in the hard-to-all condition compared to the easy-to-all condition. Overall, it seems unwise to attach much theoretical weight to an interaction where none of the simple effects are significant. However, it may be possible to pursue the issue further in future research.

Experiment 4

Spiering and Ashby (2008) reported a single experiment in which they found that initial training on difficult items facilitated performance with subsequent moderate-difficulty items, relative to initial training with easy items. In the current paper, we have reported three experiments in which this effect was not found. Indeed, these experiments all found evidence for the null hypothesis that there was no difference in performance between the two conditions.

Collectively, these three experiments indicate that the original effect in Block 2 reported by Spiering and Ashby (2008) was likely to have been a false positive (Type I error). However, a critic might potentially observe that our attempts to replicate involved only the first two blocks of Spiering and Ashby's four-block experiment. In particular, the final block of their experiment (Block 4) was used by them to support the claim that initial training on difficult items is advantageous. In Experiments 1–3, we chose not to run Blocks 3–4, because the data they generate is hard to interpret (see Introduction). Nevertheless, one could argue that it is of some interest whether the results of the second half of their experiment are real but hard to interpret, or are, alternatively, also the product of a Type I error. Hence, in this final experiment, we performed a direct replication of the full four-block design of Experiment 1 from Spiering and Ashby.

Method

Participants

The participants were 55 undergraduate psychology students recruited from the University of Exeter participation pool. They were randomly assigned to either the easy-to-hard (N=27) or hard-to-easy (N=28) condition. They received research credit in exchange for their participation.

Category structure and stimuli

The stimuli were sine-wave gratings displayed on a grey background that were identical to those used in Spiering and Ashby (2008) and Experiment 1. The stimuli used are shown in Figure 1.

Procedure

Participants were tested in individual testing booths and asked to focus on accuracy of responding. The experimental procedure was identical to that in Experiment 1, however two extra training blocks were added. There were 4 blocks of 150 trials each, resulting in a total of 600 trials. In the three training blocks, each of the stimuli were presented in a random order 15 times. The order in which the blocks were presented depended on the condition to which the participants were assigned. In the easy-to-hard condition, participants were shown only the easy stimuli, far from the category boundary in Block 1, the stimuli of moderate difficulty in Block 2 and the hard stimuli, close to the category boundary, in Block 3. In the hard-to-easy condition, the training blocks were shown to participants in the opposite order. Block 4 in both conditions showed all the stimuli in a random order, 5 times each, with feedback. This experiment was identical to that reported by Spiering and Ashby (2008).

After completing the experiment, participants were asked to complete the strategy questionnaire.

Analysis

In addition to the Bayesian analyses conducted on Blocks 1 and 2 in the experiments above, here it is also necessary to look at Block 4. In Block 4, we assumed a two-tailed normal distribution with a predicted mean difference of -0.15, and predicted standard deviation of 0.075. These values were estimated from the results presented in Spiering and Ashby (2008).

Results

The trial-level raw data are available at `www.willslab.org.uk/ply75` with md5 checksum `f3803f34d292f58eae0a7686ce784277`.

The average accuracy for each block across the experiment is shown in Figure 6. Three participants were excluded from the hard-to-easy condition because they scored below 0.3 for the majority of the experiment.[6] This resulted in 27 participants in the easy-to-hard condition and 25 in the hard-to-easy condition.

For this experiment, the Bayes Factors were calculated using the same technique and prior as described in Experiment 1 for the first two blocks. Here, for Block 1, the

[6]One participant throughout the experiment and two in Blocks 2 to 4

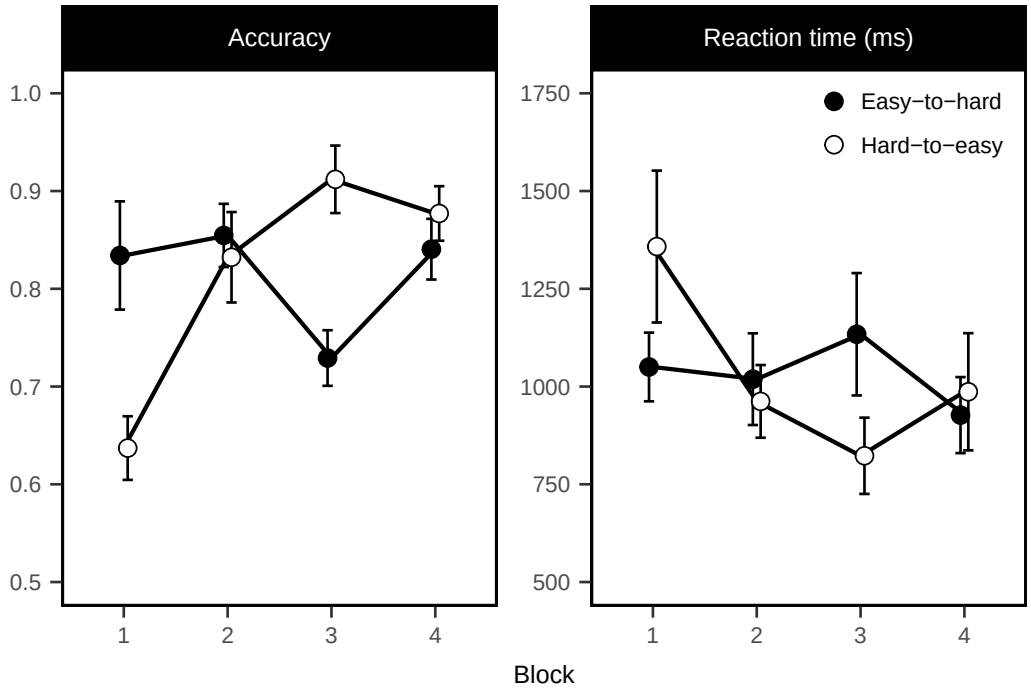

Figure 6. The average accuracy and reaction time for each block in each condition in Experiment 4. Error bars are difference-adjusted between-subject 95% confidence intervals (Baguley, 2012).

sample mean difference was 0.184, with a sample standard error of 0.045. For Block 2, the sample mean difference was 0.009, with a sample standard error of 0.038. The Bayes Factor for Block 4 was calculated using the prior defined in the Method section. The sample mean difference for Block 4 was -0.049, with a sample standard error of the difference of 0.029.

Following Spiering and Ashby (2008), we compared the mean differences between the conditions at both Block 2 and Block 4. These blocks are where the participants in both conditions saw the same stimuli (moderate-difficulty stimuli in Block 2, and all stimuli in Block 4), and so their performance can be compared fairly.

During Block 1, performance on the easy stimuli was better than performance on the hard stimuli, $t(50) = 4.09$, $d = 1.13$, $p < .001$, $BF = 1356$. However, this initial difference in accuracy did not produce a significant effect on learning performance in Block 2, $t(50) = 0.56$, $d = 0.15$, $p = .581$. Indeed there was substantial evidence for the null as the Bayes Factor was below a third, $BF = 0.09$. The Bayes Factor remains in favour of the null ($BF < 1/3$) even if the predicted mean difference is reduced by a half to -0.070, $SD_{\text{diff}} = 0.035$.

This pattern of results is also supported by the reaction time data. During Block 1, performance on the hard stimuli was slower than performance on the easy stimuli, $t(50) = 2.96$, $d = 0.82$, $p = .005$. However, in Block 2 the difference was not significant, $t(50) = 0.56$, $d = 0.15$, $p = .581$.

The data in the first two blocks are consistent with the findings of Experiments 1-3. However, to ascertain whether Spiering and Ashby (2008) was an example of a Type I error, it is also important to see whether there was any difference in performance at Block 4. In Block 4, performance in the hard-to-easy condition was numerically higher than performance in the easy-to-hard condition. This difference was not significant as

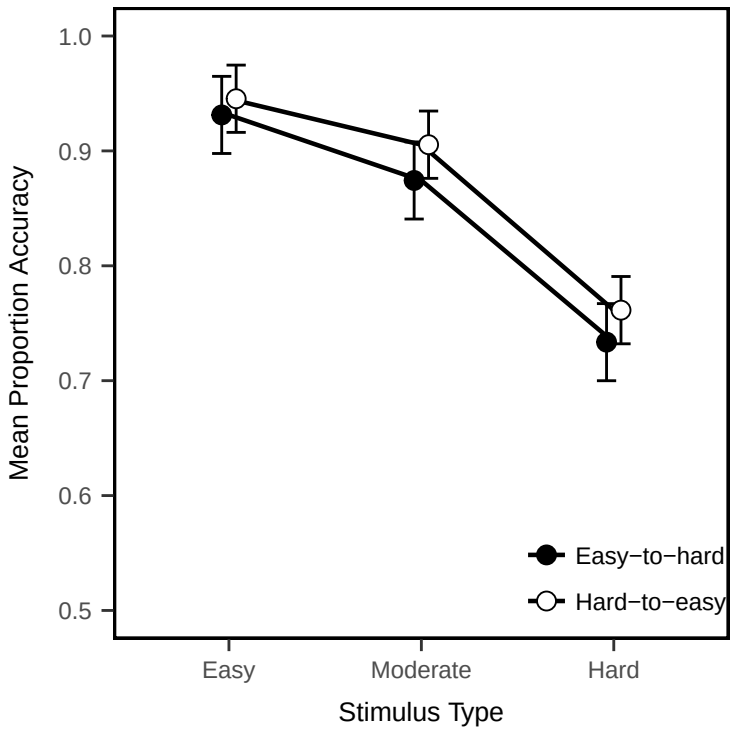

Figure 7. Average accuracy for each condition for each stimulus type in the Test (Block 4) phase of the experiment. Error bars are 95% difference-adjusted between-subject confidence intervals (Baguley, 2012).

a two-tailed test, $t(50) = 1.71$, $d = 0.48$, $p = .093$, but was significant one-tailed. One could potentially argue that a one-tailed test is merited, given the theoretical predictions and previous results of Spiering and Ashby, in which case one would conclude that we successfully replicated their Block 4 effect. On the other hand, the Bayes Factor was 0.71, indicating that it would be unwise to update one's beliefs about the veracity of Spiering and Ashby's Block 4 result on the basis of the current experiment.

A further analysis was conducted of Block 4. In Block 4, it is possible to compare accuracy for the different stimulus types in each condition as in Spiering and Ashby (2008). Such a comparison is not possible in Block 2 of the current experiment, as it contains only one type of stimulus (moderate-difficulty stimuli), and is not meaningful in Blocks 1 or 3, as people in different conditions saw different stimuli in these blocks.

The results of this additional analysis of Block 4 are shown in Figure 7. We conducted an ANOVA between condition and stimulus type on the Block 4 data. Here, the relevant contrasts are Huynh-Feldt corrected as Mauchly's test of sphericity was significant, $W = 0.77$, $p = .002$. The main effect of difficulty was statistically significant, $F(2, 100) = 143.19$, $\eta^2_G = 0.34$, $p < .001$. As one might expect, the easy stimuli were categorised with higher accuracy than the difficult stimuli. The main effect of condition reached significance only as a one-tailed test, $F(1, 50) = 2.94$, $\eta^2_G = 0.05$, $p = .093$. The interaction term did not approach significance, $F(2, 100) = 0.30$, $\eta^2_G = 0.00$, $p = .743$.

Discussion

The first half of Experiment 4 confirmed the results found in Experiments 1, 2 and 3: although there is a significant difference in performance between conditions in Block 1,

this effect disappears by Block 2. The results of the second half of Experiment 4, in particular Block 4, are best described as equivocal, due to a Bayes Factor close to unity. Nevertheless, given that the result was significant in the same direction as Spiering and Ashby observed if a one-tailed test is used, it may also be instructive to consider the interpretation of these second-half results under the assumption that they are real.

On this basis, interpreting this Block 4 effect is difficult, because the conditions vary on the stimuli seen in Block 1, the stimuli seen in Block 3 and the order of stimulus difficulty over time. However, as there was evidence for an absence of a difference in performance between the conditions in Block 2, the Block 4 effect appears more likely driven by the stimuli that the participants saw in Block 3, rather than the stimuli they saw in Block 1. Under this assumption, Block 4 of the current experiment is an example of transfer along a continuum: training on the easy stimuli (for those in the hard-to-easy condition) in Block 3 results in better performance in Block 4 than training on the hard stimuli (for those in the easy-to-hard condition). In other words, to the extent that the Block 4 result in Experiment 4 is real, it is more consistent with the well-established phenomenon of TAC, than with the approximately opposite interpretation offered by Spiering and Ashby (2008).

Strategy analyses

Model-based strategy analysis

The evidence above indicates that the Block 2 effect reported in Experiment 1 of Spiering and Ashby (2008) was a false positive and that the Block 4 effect, if genuine, is arguably more consistent with TAC than with the predictions of the dual-system model COVIS. However, a proponent of COVIS might raise the following objection: it is possible that category learning is mediated by dual systems of learning, but that the majority of participants in these particular experiments did not switch to the optimum Procedural System for the category structure (for whatever reason). If this was the case, failing to find a difference between conditions would be predicted by the COVIS model.

To overcome this sort of objection, experimental work within the COVIS literature uses a model-based analysis to determine which strategy the participant is using (e.g., Maddox & Ashby, 1993). The model-based analysis determines which of a set of experimenter-selected decision-bound models best describes the pattern of responding for each participant (Maddox & Ashby, 1993). Details of this analysis can be found in the Appendix; Table 1 summarises the main results. Specifically, Table 1 shows that, in Blocks 2 and 4, the probability of the best-fitting model being the optimum model for the category structure, is high in every experiment. According to the logic of the analysis (as used in the COVIS literature), Table 1 indicates that our participants are using the Procedural System to learn this category structure. Therefore, the failure to find evidence of initial training on final performance cannot be attributed to participants never being able to switch to the Procedural system. So, this particular objection from a COVIS perspective turns out to be unfounded.

Further inspection of Table 1 reveals some aspects of this analysis that seem counter to COVIS's predictions of what should happen in these experiments. Specifically, in Block 2 of Experiments 2, 3, and 4, the evidence for a switch to the Procedural System is greater in the easy-first condition than the hard-first condition. This is opposite to the predictions of the COVIS account. Specifically, COVIS predicts that

Table 1. The normalised probability that the optimum diagonal strategy is preferred over rule-based strategies.

Condition	Experiment			
	1	2	3	4
Easy first				
Block 1	0.34	0.29	0.21	0.50
Block 2	0.82	0.99	0.85	0.77
Block 3	-	-	-	0.66
Block 4	-	-	-	0.77
Hard first				
Block 1	0.64	0.71	0.59	0.44
Block 2	0.69	0.99	0.78	0.73
Block 3	-	-	-	0.51
Block 4	-	-	-	0.88

the hard-first training should increase the proportion of participants found to be using the Procedural System (opposite to what is observed in this analysis). Block 2 of Experiment 1 is at ceiling on this measure and so uninformative in this regard. Only Block 4 of Experiment 4 shows the COVIS-predicted direction of effect although this, like the behavioural result itself, is hard to interpret due the inherent primacy-recency confound of the second half of this design (see Introduction).

Looking further at the data from Table 1, it may also be tempting to draw conclusions about Block 1 (and Block 3 of Experiment 4). For instance, looking at the entries for Block 1 in Table 1, you can see that the probabilities that the optimum (diagonal) strategy is used are higher in the hard-first condition than the easy-first condition. Some might argue that this pattern of results supports the predictions of COVIS: that participants in the hard first conditions switch faster to the Procedural System and can thus implement the optimum diagonal strategy faster. Unfortunately, the model-based analysis cannot be used in this way.

This comparison is problematic because the results of the strategy analysis are conditional on the category structure it is applied to. Here, the issue lies in the fact that for the easy stimuli the unidimensional strategy model would result in similar performance to the optimum diagonal strategy model. Whereas for the hard stimuli, the unidimensional strategy would score much lower than the optimum diagonal strategy. This issue is exaggerated as the unidimensional strategy model has two parameters whereas the diagonal general linear classifier strategy model has three. Therefore, if the fit is similar between these two models the model-fitting procedure typically used in this procedure will always favour the simpler (lower parameter) model. This biases the analysis towards finding more unidimensional strategies in Block 1 of the initially easy conditions than the initially hard conditions (for a more thorough discussion of these issues see Edmunds, Milton, & Wills, 2017; Pitt, Myung, & Zhang, 2002). More

generally, the sort of strategy modelling commonly included in COVIS-inspired papers is likely to be inconclusive due to a range of methodological issues (Donkin, Newell, Kalish, Dunn, & Nosofsky, 2015; Edmunds et al., 2017). Such modelling should thus should be interpreted with caution. We have included this kind of modelling in the current paper because, despite its limitations, it plays a key role in the logic of most COVIS-inspired experiments.

In summary, it is only in Blocks 2 and 4, where participants in both conditions see the same stimuli, that the results of such analyses can straightforwardly be compared. The results for Block 2 and 4 show a high probability that the optimum (diagonal) decision-bound model is the best-fitting of those tested, supporting the applicability of the COVIS account. However, the effect of experiment condition on these Block 2 and 4 probabilities is largely contrary to the predictions of COVIS.

Analysis of participants' strategy reports

In research supporting the COVIS model, participants that are found to be using the optimum (diagonal) strategy by the above model-based analyses are assumed to be responding on the basis of implicit knowledge (Smith et al., 2015). However, recent evidence has found that the proportion of participants identified by model-based techniques as using "implicit" strategies varies depending on the details of the analysis procedure used. For instance, if the range of rule-based strategies included within the model-based analysis procedure is increased, the proportion of participants classified as "implicit" (diagonal) responders goes down substantially (Donkin et al., 2015). Therefore, it seems wise to further investigate the assumption that an optimum (diagonal) strategy as indicated by model-based analysis means that the participant is responding on the basis of implicit knowledge.

To conduct this investigation, which follows an earlier investigation of a similar type reported by Edmunds et al. (2015), we asked participants to describe the strategies they used. This type of awareness task has also been previously conducted in other investigations of implicit learning and, if anything, over-estimates the numbers of implicit responders (e.g., Konstantinidis & Shanks, 2014; Newell & Shanks, 2014; Shanks & St. John, 1994; Yeates, Jones, Wills, Aitken, & McLaren, 2013). As the COVIS model assumes that the diagonal strategies are learned implicitly (Ashby et al., 1998), it predicts that the majority of participants in our experiments would not be able to report any clear strategy (because model-based analysis identifies them as using a diagonal strategy). Further, as participants are predicted to switch to the implicit system faster in the hard-to-easy conditions, COVIS predicts that fewer participants should be able to report a strategy in the hard-first conditions than in the easy-first conditions.

An alternative possibility, not particularly consistent with COVIS, is that participants use strategies that combine information across both dimensions, but that these strategies are not implicit. In other studies of classification behaviour, outside the COVIS framework, participants commonly report these kinds of strategy (e.g., Wills, Milton, Longmore, Hester, & Robinson, 2013).

The strategy-report questionnaires that we administered in the current experiments were independently coded by the first author (CERE) and a student volunteer (GW). First, each verbal report was examined to determine whether the participant had reported a clear categorisation strategy or not. Second, the available strategy descriptions were sorted into the groups specified below — all descriptions were classifiable

Table 2. Summary of the inter-rater reliability statistics for judging whether the participant had reported a strategy, and for the type of strategy identified. Also listed are the number of participants in each condition of each experiment that did not report a strategy.

Experiment	Presence of Strategy		Type of Strategy		no strategy N	
	κ	p-value	κ	p-value	Easy first	Hard first
Experiment 1	0.84	$< .001$	0.49	$< .001$	1	2
Experiment 3	0.66	$< .001$	0.70	$< .001$	1	-
Experiment 4	1.00	$< .001$	0.89	$< .001$	2	3

into one of these groups. The inter-rater reliabilities for these initial codings are reported in Table 2. Then, any discrepancies between raters were easily resolved through discussion with reference to the strategy descriptions below. Across the three experiments for which we had data (due to an oversight, no questionnaires were administered in Experiment 2), the following types of strategy were identified:

Participants were classified as using a *complex rule* if they described a rule using both stimulus dimensions in a complicated fashion. Example strategies include rule-plus-exception strategies such as "upright stimuli were in Category A and flat stimuli in Category B. However, if the stimulus was upright and had very few bars it was in Category B" or sequential unidimensional rules such as "upright stimuli were in Category A and flat stimuli were in Category B. For stimuli at 45 degrees, it was in Category A if it had less than three bars and Category B if it had more than three bars."

Participants were classified as using a *conjunction* rule if they used both stimulus dimensions and described categorising stimuli using a logical conjunction rule such as "upright stimuli with lots of lines were in Category A, otherwise they were in Category B."

Participants were classified as using a *two-dimensional* rule if they described using both stimulus dimensions but with descriptions that were too unclear to be assigned to more specific categories.

Participants were classified as using a *unidimensional* rule if they described categorising stimuli based solely on either bar frequency or stimulus orientation.

In addition, a few participants described elements of the experimental setup, such as which buttons to press, rather than their sorting strategy; these participants were considered to have not reported a strategy.

We also created an *overall similarity* classification. If any participant had described attempting to make the stimulus dimensions commensurable, such as "Stimuli for which the line was longer than it was upright should be assigned to category A", or if they had said anything that could have reasonably been interpreted as a statement that they based their classification on overall similarity, they would have been placed in this group. In practice, across all three experiments, no participants made reports of this type. Further, no participants reported using an implicit strategy, such as "I went with my gut" or any similar or comparable statement.

Table 3. The proportion of participants that reported using each strategy type.

Condition	Verbal reports			
	2D	CJ	Complex	UD
Experiment 1				
Easy-to-moderate	0.06	0.19	0.69	0.06
Hard-to-moderate	0.11	0.22	0.56	0.11
Experiment 3				
Easy-to-all	0.12	0.18	0.47	0.24
Hard-to-all	0.05	0.45	0.50	0.00
Experiment 4				
Easy-to-hard	0.04	0.12	0.48	0.36
Hard-to-easy	0.00	0.18	0.68	0.14

Strategies: 2D = rule using both dimensions, CJ = Conjunction, UD = Unidimensional.

The number of participants not reporting a strategy in each condition of each experiment are displayed in Table 2 — as can be observed, the vast majority of participants reported a classifiable sorting strategy. Further, as can be seen in Table 3, participants reported a range of explicit rules, with complex rules being the most common in all conditions of all experiments. This is inconsistent with COVIS, which, given the results of the model-based analyses, predicts that most participants should have learned this category structure structure implicitly and thus, at least under normal definitions of the term "implicit", should not be able to report a strategy. Even if one considers the two-dimensional rule-type, due to its vagueness, as a failed attempt to describe an implicit strategy, it still remains the case that the vast majority of participants in all conditions of all experiments in the current paper described a clearly-expressed rule when asked how they had classified the stimuli.

Of course, it is possible to consider these kinds of subjective reports as epiphenomenal, and attribute classification behaviour to implicit processes that operate independently of whatever mental process leads to these reports. Further research is required to resolve this issue definitively, but the idea that information-integration category structures are learned implicitly does not seem to be required by the currently available evidence.

General discussion

In the current paper, we examined how initial training difficulty impacts final category learning performance. The literature highlighted two possibilities. First, the well-established phenomenon of Transfer Along a Continuum (Lawrence, 1952) leads to the prediction that participants initially trained on the easy stimuli would perform better at test than those initially trained on hard stimuli. In contrast, Spiering and

Ashby (2008) reported an experiment in which initial training on a hard discrimination resulted in better performance than initial training on an easy discrimination. This result, approximately opposite to TAC, can be predicted by the COVIS model of category learning. Specifically, Spiering and Ashby argued that participants in the easy-to-hard condition could perform very well on the easy stimuli by using a suboptimum unidimensional strategy and so would delay swapping from the Explicit System to the optimum Procedural System. In contrast, participants who were first shown the difficult stimuli would swap to the optimum Procedural System much sooner as it would be clear that rule-based approaches were not working.

We reported four experiments, all of which failed to support the conclusions of Spiering and Ashby (2008). In Experiment 1, we re-examined the first two blocks of Spiering and Ashby's experiment; the later parts (Blocks 3 and 4) of their experiment were difficult to interpret due to the confound discussed in the Introduction. In contrast to Spiering and Ashby, we found substantial Bayesian evidence for the null hypothesis: the type of initial training had no effect on final performance. In Experiment 2, we changed the feedback and stimuli to rule out the possibility that they were confusing the participants. Here, we again found evidence for the null hypothesis. In Experiment 3, we examined the possibility that Experiments 1 and 2 were subject to a ceiling effect. To do this, we imposed time pressure to reduce overall performance. Once again, we failed to find a differential effect of initial training, but did find substantial evidence for the null hypothesis. Finally, in Experiment 4, we replicated the entirety of Spiering and Ashby's experiment. Here, in Block 2, we once again found evidence for the null hypothesis. However, in Block 4, if we used a one-tailed test, participants in the hard-to-easy training condition performed significantly better than the participants in the easy-to-hard training condition. Bayesian analysis indicates this result is equivocal (BF close to unity). Nevertheless, if one assumes the Block 4 effect is real, the evidence for the absence of an effect in Block 2 means this final result is more compatible with TAC than with the approximately opposite conclusions of Spiering and Ashby.

So, all in all, our results indicate that the effect found in Spiering and Ashby (2008) was likely a false positive. In other words, a sufficient explanation for the difference between their single study and the several reported here is an unfortunate initial sample of the population on the part of Spiering and Ashby. This is a known hazard of running a regular-sized study a single time.

Is there ever a difficult-first benefit in category learning?

A comprehensive answer to this question is beyond the scope of the current article, or any other relatively brief empirical report. Nevertheless, it may be worth considering one previous publication, discussed at length by Spiering and Ashby, which also appears to show a difficult-first benefit in training order in category learning (Lee, MacGregor, Bavelas, Mirlin, & Lam, 1988).

Lee et al.'s (1988) paper differs from other work in this area in a number of important respects. Perhaps most strikingly, Lee et al. use a cascade of participants to define stimulus difficulty empirically. So, participant 1 (P1) gets the stimuli in a random order. P2 gets the stimuli P1 got wrong first, followed by the stimuli P1 got right. This then iterates to P3. Lee et al. found that, in general, P3 learns more quickly than P2, who learns more quickly than P1. This is different to other experiments in this area, which all use some experimenter-defined notion of stimulus difficulty (e.g.,

distance from the optimal bound in stimulus space). It seems a reasonable conjecture that the use of a participant cascade in this manner might lead to an ordering of stimuli that was well correlated with an assessment of difficulty based on physical stimulus properties. However, Lee et al. provide no information that could be used to confirm or reject this conjecture.

Nevertheless, let us assume for a moment that the conjecture is correct. On this basis, the results of Lee et al. (1988) seem problematic for a COVIS account, because the same difficult-first benefit is observed for both rule-defined category structures (Exp. 1, 4) and category structures likely to be information-integration from a COVIS perspective (e.g. male vs. female handwriting in Exp. 3). COVIS predicts a difficult-first benefit should only be observed with an information-integration category structure (Spiering & Ashby, 2008).

Spiering and Ashby (2008) counter-argue that the presence of a difficult-first benefit in a rule-based category in Lee et al. (1988) is not a problem for COVIS, because Lee et al. failed to include a transfer test (i.e., an equivalent to Blocks 2 and 4 in the current experiments, where all participants are presented with the same set of stimuli in a random order). To illustrate the problem this absence of a common transfer test causes, imagine that the participant's knowledge of the category structure is better at the end of the experiment than the beginning, but this relationship between knowledge and experience is unaffected by the order in which the items are presented. Under these conditions, it can still be the case that the difficult-first participant scores better overall. For example, when a participant has zero knowledge, the item difficulty is irrelevant as all responses will be guesses, while when the participant has good but imperfect knowledge their accuracy will be higher for easy than for difficult items. So, putting difficult items first can lead to higher mean accuracy for reasons other than promoting better learning.

In pointing this out, Spiering and Ashby (2008) make an insightful critique of Lee et al. (1988). However, the critique potentially applies to all experiments in Lee et al. and hence also undermines the claim that there have been previous demonstrations of a difficult-first benefit in an information-integration category structure. In summary, Lee et al. provides no compelling evidence that difficult-first training confers a benefit to category learning.

Spiering and Ashby cite two further articles that they seem to imply support a difficult-first benefit in procedures other than category learning (Ahissar & Hochstein, 1997; Doane, Sohn, & Schreiber, 1999). However, the relation between these studies and the others considered in the current article is remote at best, and they provide no compelling reason to revise our assessment that the results of Spiering and Ashby are most likely a false positive.

The role of counter-intuitive feedback

Other than simply being a case of a Type I error, another possible explanation for why Spiering and Ashby (2008) found the result that they did is that they may have failed to exclude mis-learners (their paper is unclear on this point). Mis-learners are those participants whose performance was markedly different from chance, but in the wrong direction, i.e. they scored around 10% rather than 90%. These participants have obviously learned the features of the category structure, but mistook the response key for each category structure. This is to be compared with non-learners who score around chance (here 50%), who have not learned the features of the category structure.

Mis-learning could also have been a problem for Spiering and Ashby's experiment because of their counter-intuitive choice of feedback signals: their high tone indicated an incorrect response, whereas their low tone indicated a correct response. When we used this feedback (Experiments 1 and 4) several participants mis-learned the category structure. When we used more typical feedback signals, there were no mis-learners.

The mis-learning participants are important because the conclusions one draws from the experiment depends on whether or not they are included in the analyses. In Experiment 1, three mis-learning participants were excluded from the easy-to-moderate condition. If they are instead included, the pattern of results is similar to Spiering and Ashby (2008, although it fails to reach significance): hard initial training results in better Block 2 performance than easy initial training. In Experiment 4, three mis-learning participants were removed from the hard-to-easy condition. In this case, any difference between conditions in Block 4 disappears if these mis-learners are included. Therefore, the impact of including mis-learners depends on which condition they were in. This raises possibility that the original effect was due to several participants in the easy-to-hard condition mis-learning the category structure and not being excluded.

Not only do these experiments indicate that the conclusions we can draw from them are sensitive to mis-learners, they also highlight the importance of applying a learning criterion to experiments within the COVIS literature. This observation is consistent with other critiques of studies within the COVIS literature. For example, Zeithamova and Maddox (2006) looked at the effect of concurrent load on category learning. As predicted by the COVIS model, Zeithamova and Maddox found that concurrent load negatively impacted rule-based but not information-integration category learning. However, when Newell et al. (2010) re-examined these experiments, they found that the conclusions they could draw were dependent on the inclusion of non-learners. When non-learners were included in these experiments, the data were consistent with the COVIS model. However, when they were excluded the experiment failed to find evidence of a dual-system model.

What does this mean for the COVIS model?

The current work adds to the growing literature that weakens the evidential support for the COVIS model (such as Carpenter, Wills, Benattayallah, & Milton, 2016; Edmunds et al., 2015; Newell et al., 2010; Nosofsky & Kruschke, 2002; Nosofsky, Stanton, & Zaki, 2005; Stanton & Nosofsky, 2007, 2013; Zaki & Kleinschmidt, 2014). These studies have found that much of the evidence argued to support the dual-process COVIS model is amenable to alternative, often single-system, explanations. On the face of it, a single-system, rule-based, account also seems sufficient to explain the current results. In these studies, we asked participants to describe the strategy that they used to complete the category-learning task. The vast majority of participants reported a specific rule-based strategy.

An alternative, COVIS-consistent, interpretation of the participants' strategy reports is that they were failed attempts to verbalize an nonverbalizable implicit classification process. If one accepts this possibility, we think one must further ask what evidence supports this interpretation over the apparently more straightforward possibility that participants were reporting the rules they used?

There seem to be two sorts of answer here. The first is that the model-based analysis employed in these studies reveals the underlying implicit process (by showing that a diagonal strategy best fits the participants' responses). We doubt this conclusion,

mainly due to work conducted by ourselves and others, which has revealed a number of methodological shortcomings in this type of model-based analysis (Donkin et al., 2015; Edmunds et al., 2017). The second sort of answer is that there are many other papers that support the COVIS theory, increasing the likelihood that the above COVIS-inspired interpretation is the correct one here, too. However, as we have just noted, such COVIS-supporting results as have been independently investigated turn out to be less clear than they first appeared.

Relationship to TAC

Previous studies have found evidence of TAC in humans (Church et al., 2013; Hornsby & Love, 2014; Suret & McLaren, 2003), so why was it so difficult to find here? One possibility consistent with the theoretical account given by McLaren and Mackintosh (2002) is that the appearance of TAC in humans critically depends on the similarity of the stimuli. In our experiments, compared to previous demonstrations, even the stimuli most similar to each other appear obviously different, especially those in Category A (see Figure 1). Therefore, perhaps we did not find TAC in our experiments because the stimuli were not sufficiently similar to each other.

Another possible explanation might relate to the pre-experimental salience of the stimulus dimensions, and their relevance to the category structure to be learned. The stimuli used in previous demonstrations of TAC in humans, such as morphed faces (Suret & McLaren, 2003), or mammograms (Hornsby & Love, 2014), have dimensions of variation that may not be immediatley apparent to participants. In contrast, in Spiering and Ashby (2008) and in the current experiments, there are two pre-experimentally salient dimensions (bar frequency and orientation), neither of which are individually good predictors of category membership. For such stimuli, participants might require quite extensive exposure to overcome the initial saliences of these imperfectly-predictive stimulus dimensions, and focus on the critical discrimination dimension (i.e. the minor diagonal in stimulus space), which provides the basis for a TAC effect.

Conclusion

The current work challenges Spiering and Ashby's claim, made on the basis of a single experiment, that it is sometimes best to start training with the most difficult items. All-in-all, the current work illustrates the dangers of making striking, novel, claims on the basis of any one experiment.

References

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, *387*, 401–406.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 65–87). Cambridge University Press.

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*, 158–175.

Baguley, T., & Kaye, D. (2010). Book review: Understanding psychology as a science: An introduction to scientific and statistical inference. *British Journal of Mathematical and Statistical Psychology*, *63*, 695–698.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Carpenter, K. L., Wills, A. J., Benattayallah, A., & Milton, F. (2016). A comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*, *37*, 3557–3574.

Church, B. A., Mercado, E., Wisniewski, M. G., & Liu, E. H. (2013). Temporal Dynamics in auditory perceptual learning: Impact of sequencing and incidental learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 270–276.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Doane, S., Sohn, Y., & Schreiber, B. (1999). The role of processing strategies in the acquisition and transfer of cognitive skill. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 1390–1410.

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., & Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: A case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 933–948.

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. *Quarterly Journal of Experimental Psychology*, *68*, 1203–1222.

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2017). Due process in dual-process: Model recovery simulations of General Recognition Theory analysis. *Submitted to Cognitive Science*.

Ell, S. W., & Ashby, F. G. (2006). The effects of category overlap on information-integration and rule-based category learning. *Perception & Psychophysics*, *68*, 1013–1026.

Filoteo, J. V., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. *Psychological Science*, *21*, 415–423.

Hornsby, A. N., & Love, B. C. (2014). Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*, *3*, 72–76.

Jeffreys, H. (1961). *The Theory of Probability* (3rd ed.). Oxford: Oxford University Press.

Konstantinidis, E., & Shanks, D. R. (2014). Don't bet on it! Wagering as a measure of awareness in decision making under uncertainty. *Journal of Experimental Psychology: General*, *143*, 2111–2134.

Lawrence, D. H. (1952). The transfer of a discrimination along a continuum. *Journal of Comparative and Physiological Psychology*, *45*, 511–516.

Lee, E. S., MacGregor, J. N., Bavelas, A., Mirlin, L., & Lam, N. (1988). The effects of error transformations on classification performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 66–74.

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin*, *142*, 1111–1140.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*, 276–298.

Mackintosh, N. J., & Little, L. (1970). An analysis of transfer along a continuum. *Canadian Journal of Psychology*, *24*, 362–369.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*, 49–70.

Matsuki, K. (2014). *GRT: General Recognition Theory (0.2).* https://cran.r-project.org/package=grt.

McLaren, I. P. L., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning and Behavior*, *30*, 177–200.

Newell, B. R., Dunn, J. C., & Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition*, *38*, 563–581.

Newell, B. R., Moore, C., Wills, A. J., & Milton, F. (2013). Reinstating the frontal lobes? Having more time to think improves 'implicit' perceptual categorization. a comment on Filoteo, Lauritzen and Maddox (2010). *Psychological Science*, *24*, 386–389.

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *37*(1), 1–19.

Nosofsky, R. M., & Kruschke, J. K. (2002). Single-system models and interference in category learning: Commentary on Waldron and Ashby (2001). *Psychonomic Bulletin & Review*, *9*, 169–174.

Nosofsky, R. M., Stanton, R. D., & Zaki, S. R. (2005). Procedural interference in perceptual classification: implicit learning or cognitive complexity? *Memory & Cognition*, *33*, 1256–1271.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

R Core Team. (2017). *R: A language and environment for statistical computing.* https://www.r-project.org/.

Roeder, J. L., & Ashby, F. G. (2016). What is automatized during perceptual categorization? *Cognition*, *154*, 22–33.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of Dissociable Human Learning-Systems. *Behavioral and Brain Sciences*, *17*, 367–395.

Smith, J. D., Zakrzewski, A. C., Herberger, E. R., Boomer, J., Roeder, J. L., Ashby, F. G., & Church, B. A. (2015). The time course of explicit and implicit categorization. *Attention, Perception, & Psychophysics*, *77*, 2476–2490.

Spiering, B. J., & Ashby, F. G. (2008). Initial training with difficult items facilitates information-integration but not rule-based category learning. *Psychological Science*, *19*, 1169–1177.

Stanton, R. D., & Nosofsky, R. M. (2007). Feedback interference and dissociations of classification: evidence against the multiple-learning-systems hypothesis. *Memory & Cognition*, *35*, 1747–1758.

Stanton, R. D., & Nosofsky, R. M. (2013). Category number impacts rule-based and information-integration category learning: A reassessment of evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1174–1191.

Suret, M., & McLaren, I. P. L. (2003). Representation and discrimination on an artificial dimension. *Quarterly Journal of Experimental Psychology*, *56B*, 30–42.

Sutherland, N. S., & Mackintosh, N. J. (1971). *Mechanisms of animal discrimination learning.* New York: Academic Press.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196.

Wills, A. J., Milton, F., Longmore, C. A., Hester, S., & Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification? *Quarterly Journal of Experimental Psychology*, *66*, 299–318.

Yeates, F., Jones, F., Wills, A., Aitken, M., & McLaren, I. (2013). Implicit learning: A demonstration and a revision of a novel SRT paradigm. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (Vol. 1, pp. 3829–3834). Austin, TX: Cognitive Science Society.

Zaki, S. R., & Kleinschmidt, D. F. (2014). Procedural memory effects in categorization: Evidence for multiple systems or task complexity? *Memory & Cognition*, *42*, 508–524.

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34*, 387–398.

Appendix

Although the kind of model-based analysis described in this Appendix is ubiquitous in the experimental COVIS literature, the types of strategy models included, and their precise specifications, often vary between papers. In order to facilitate comparison with the analysis presented in Spiering and Ashby (2008), here we use the same set of models as employed in their paper.

The set of models considered by Spiering and Ashby (2008) were of three main types: rule-based, information-integration and random models. Within the COVIS framework, the unidimensional and conjunction models are considered to represent explicit, rule-based strategies, while the diagonal general linear classifier (GLC) strategy is considered to represent an implicit, information-integration strategy. The strategy models used in this analysis were specified as follows:

The *unidimensional* models assume that the participant determines a criterion along one of the stimulus dimensions, either orientation or length (or bar width, depending on the stimulus type). They then make a decision about the category membership of each stimulus by comparing the appropriate stimulus attribute with the criterion value. As an example, for length, this corresponds to a rule of the type: "Assign to Category A if the stimulus is long, or Category B if short." The unidimensional models have two parameters: the value of the criterion and the variance of internal (criterial and perceptual) noise.

The *conjunction* model assumes that the participants make two judgements, one for each stimulus dimension, and then combine these to make a judgement about category membership. The conjunction rule in the current analysis was of the type: "Assign to Category A if the stimulus is short and upright, otherwise assign to Category B." The conjunction model had three parameters: the two criterion values and internal noise.

The *general linear classifier (GLC)* model assumes that the decision boundary between the categories can be described by a straight line that can vary in gradient and intercept. The unidimensional models are therefore special cases of the GLC model. The GLC model has three parameters: the intercept and slope of the decision bound, plus noise.

There are two random models that assume that participants are responding randomly. The *random* model assumes that participants have no preference for either category: it has no parameters. The *random bias* model assumes that participants respond randomly but prefer one category over the other. It has one parameter that represents the amount of bias.

For each participant, the fit of each of these models was calculated using the Bayesian Information Criterion (BIC; Schwarz, 1978)

$$BIC = r \ln N - 2 \ln L \tag{1}$$

where r is the number of parameters in the model, N is the sample size and L is the likelihood of the model given the data. The results from this analysis, which was performed using the `grt` package in the R environment (Matsuki, 2014), are reported in Table 4. Table 4 shows that a majority of participants in the test blocks (Block 2 and 4) of each condition in each experiment were found to be using the optimum

diagonal strategy for the category structure.

Table 4. The proportion of participants that were assigned to each strategy according to the model-based strategy analysis based on the responses from each block for each experiment.

Condition	Strategies *(wBIC)*				
	GLC	CJ	UD	RND	BIAS
Experiment 1					
Easy-to-moderate					
Block 1	0.41 *(0.76)*	0.41 *(0.98)*	0.18 *(0.80)*	-	-
Block 2	0.82 *(0.96)*	-	0.18 *(0.74)*	-	-
Hard-to-moderate					
Block 1	0.60 *(0.95)*	-	0.20 *(0.81)*	0.15 *(0.59)*	0.05 *(0.63)*
Block 2	0.65 *(0.97)*	0.10 *(0.75)*	0.15 *(0.77)*	0.10 *(0.79)*	-
Experiment 2					
Easy-to-moderate					
Block 1	0.35 *(0.80)*	0.55 *(0.98)*	0.10 *(0.81)*	-	-
Block 2	1.00 *(0.99)*	-	-	-	-
Hard-to-moderate					
Block 1	0.61 *(0.98)*	0.13 *(0.65)*	0.17 *(0.64)*	0.04 *(0.87)*	0.04 *(0.31)*
Block 2	1.00 *(0.99)*	-	-	-	-
Experiment 3					
Easy-to-all					
Block 1	0.17 *(0.84)*	0.56 *(0.93)*	0.28 *(0.75)*	-	-
Block 2	0.89 *(0.93)*	-	0.11 *(0.60)*	-	-
Hard-to-all					
Block 1	0.60 *(0.88)*	-	0.30 *(0.79)*	0.10 *(0.73)*	-
Block 2	0.80 *(0.94)*	-	0.15 *(0.69)*	0.05 *(0.81)*	-
Experiment 4					
Easy-to-hard					
Block 1	0.46 *(0.91)*	0.08 *(0.64)*	0.42 *(0.74)*	0.04 *(0.89)*	-
Block 2	0.74 *(0.98)*	0.04 *(0.41)*	0.19 *(0.73)*	0.04 *(0.89)*	-
Block 3	0.63 *(0.97)*	-	0.30 *(0.80)*	0.04 *(0.90)*	0.04 *(0.74)*
Block 4	0.78 *(0.96)*	-	0.19 *(0.84)*	0.04 *(0.87)*	-
Hard-to-easy					
Block 1	0.43 *(0.88)*	-	0.29 *(0.74)*	0.21 *(0.69)*	0.07 *(0.51)*
Block 2	0.72 *(0.95)*	0.12 *(0.70)*	0.08 *(0.72)*	0.08 *(0.73)*	-
Block 3	0.48 *(0.91)*	0.12 *(0.62)*	0.36 *(0.55)*	0.04 *(0.60)*	-
Block 4	0.88 *(0.99)*	-	0.08 *(0.78)*	0.04 *(0.76)*	-

Strategies: GLC=General linear classifier, CJ=Conjunction, UD=Unidimensional, RND=Random.

Although not typically a part of the standard model-based strategy analysis used in the COVIS literature (although see Roeder & Ashby, 2016, for a Bayes Factor approach), it is also informative to look at the performance of the best-fitting model relative to the performance of the competing models. If the winning model performs

much better than its competitors, i.e. it fits better to the data, we can be more confident that this model provides the best description of the participant's behaviour from among the pre-specified alternatives. On the other hand, if the winning model performs only slightly better than the alternatives, our confidence that the winning model best describes the participant's responses should be lower. There are several cases where this might occur. For example, the participant may be swapping between strategies, applying a single strategy inconsistently with lapses in attention or even using a strategy not included within the set of models the analysis can select from (Donkin et al., 2015). Therefore, it is important to investigate the fit of the strategy models.

One principled way of evaluating the validity of the model-based analysis is by calculating Schwarz weights (Wagenmakers & Farrell, 2004). Schwarz weights ($w_i(BIC)$) are defined as the probability that model i is best, in term of minimising the BIC, given the data and the set of competing models. The average Schwarz weights for the winning models are included in Table 4. From these, it is also possible to calculate the normalised probability that the optimum diagonal strategy is preferred over rule-based strategies (i.e. conjunction and unidimensional) for each participant. From the Schwarz weights, the normalised probability that the diagonal strategy model is to be preferred over the conjunction and unidimensional strategy models is calculated using:

$$\frac{w_{GLC}(BIC)}{w_{GLC}(BIC) + w_{CJ}(BIC) + w_{UD}(BIC)} \tag{2}$$

where $w_{GLC}(BIC)$, $w_{CJ}(BIC)$ and $w_{UD}(BIC)$ are the Schwarz weights for the diagonal, conjunction and unidimensional strategy models respectively. These values are reported for each experiment in Table 1.